Supplementary Materials for

The phenotypic legacy of admixture between modern humans and Neanderthals

Corinne N. Simonti, Benjamin Vernot, Lisa Bastarache, Erwin Bottinger, David S. Carrell, Rex L. Chisholm, David R. Crosslin, Scott J. Hebbring, Gail P. Jarvik, Iftikhar J. Kullo, Rongling Li, Jyotishman Pathak, Marylyn D. Ritchie, Dan M. Roden, Shefali S. Verma, Gerard Tromp, Jeffrey D. Prato, William S. Bush, Joshua M. Akey†, Joshua C. Denny†, John A. Capra*

† Equal contribution
*Correspondence to: tony.capra@vanderbilt.edu

This PDF file includes:

Other Supplementary Materials for this manuscript include the following:

Table S2 is available online as a downloadable Excel file.

Materials and Methods

eMERGE Network Genotype and Phenotype Data

The eMERGE Cohort

The eMERGE Network is comprised of ten sites: seven with adult samples and three with pediatric samples. We used adult (18 years of age or older as of January 2015) individuals of European ancestry from the seven adult sites: Geisinger Health System, Group Health Cooperative (Washington State), Mayo Clinic, Marshfield Clinic, Mt. Sinai, Northwestern University, and Vanderbilt University. The eMERGE Phase 1 (E1) data set comprises 13,686 individuals from five of the seven sites with adult samples: Group Health Cooperative, Mayo Clinic, Marshfield Clinic, Northwestern University, and Vanderbilt University. The eMERGE Phase 2 (E2) comprises an independent set of 14,730 individuals from all seven sites.

Genotyping, Quality Control, and Imputation

The eMERGE subjects were genotyped by the Network on a range of genome-wide arrays including the Affymetrix 6.0 and the Illumina 550, 610, 660, 1M, 5M, and Omni Express chips. The eMERGE Coordinating Center at Pennsylvania State University performed genotype imputations for all samples collected as part of E1 and E2. In the v3 (December 2014) data release used here, SHAPEIT2 (*26*) and IMPUTE2 (*27*) were used to impute all autosomes to 1000 Genomes Project (release March 2012). Imputed data for all sites were then merged based on an intersection of successfully imputed SNPs between them. For different analyses, probabilities (dosages) or the most likely imputed genotypes (hard calls) were used as indicated in the text. For the hard call SNPs, the marker call rate threshold was set at 99% and info score threshold at 0.7.

Population structure was evaluated using Eigensoft 6.0 (*28*) on filtered and LD pruned data. Related individuals (one from each pair of kinship >0.125) as estimated from IBD were removed before principal components analysis (PCA). Two highly correlated regions as well as all palindromic SNPs were removed, in addition to a marker call rate filter of 99%, MAF > 10%, and LD pruning of $r^2 > 0.1$. This left 101,000 SNPs, on which 30 principal components were calculated. See the eMERGE Network methods publications and web site for more details (*29, 30*).

Phenotyping

Clinical phenotypes were derived using a prior EHR-based phenome-wide association study (PheWAS) approach, which uses established algorithms that integrate the range of diseases, signs and symptoms, causes of injury, and procedures represented by International Classification of Diseases, ninth edition (ICD-9), codes into 1,645 coherent phenotypes, such as “inflammatory bowel disease,” and its child terms “Crohn’s disease” and “ulcerative colitis” (*31*). These phenotypes and the corresponding controls (defined by lack of related codes) have seen extensive manual and computational validation across eMERGE, and they have proven successful in previous studies (*19, 32-34*). The ICD-9 based phenotype definition algorithms used produce phenotypes that enable replication of 66% of known associations in sufficiently powered (80%) association tests (*19*) and further recent unpublished work has yielded even higher replication rates.

ICD-9 code counts were extracted from the electronic health record (EHR) and converted to PheWAS code counts. From the PheWAS code count list, we used the PheWAS package (v0.9.5.1-1) (*31*) function "createPhewasTable" to generate case/control status for our individuals using a minimum code count of two unique dates of a diagnosis. We did not analyze phenotypes occurring in fewer than 20 patients.

Identification of Introgressed Neanderthal SNPs

To identify Neanderthal-introgressed variants, we first collected all variants in a set of high-confidence (FDR < 0.05) introgressed haplotypes recently identified using the S* algorithm, the Altai Neanderthal sequence, and 1KG data (*6*). This involved first calculating the statistic S*, which takes into account both divergence and linkage disequilibrium, and then refining the resulting set of candidate introgressed regions by directly comparing significant S* haplotypes to the Altai Neanderthal genome sequence. In particular, a "Neanderthal match *P* value" was calculated to quantify whether the observed matching is higher than would be expected by chance (*6*). Only haplotypes that are significant by S* and have a significant Neanderthal match *P* value are used in the analyses here. From these, we took all biallelic variants where the putatively introgressed allele matched the Altai Neanderthal sequence and was derived with respect to chimpanzee. For each haplotype, we calculated a 90% trimmed mean allele frequency. To remove variants that were unlikely to be present on the introgressed Neanderthal haplotype at the time of introgression, we restricted our variants of interest to those within 10% of the trimmed mean allele frequency of the haplotype. After these filtering steps, we required that at least four variants remained to include a haplotype and its variants. It is difficult to directly quantify how much these additional criteria decrease false positives in the original Neanderthal haplotype calls, but we anticipate that this is a considerably higher confidence set.

Genome-wide Complex Trait Analysis

Genetic Relationship Matrix (GRM) Generation

For individuals in E1 discovery set, we used all directly genotyped Neanderthal variants (on the Human660W-Quadv1_A platform) with a MAF > 1% (1,532 variants) to compute a GRM using the "make-grm" option in the GCTA program (v1.24.4) (*15, 35*). For individuals in the E2 replication set, we computed a Neanderthal GRMs using the same SNPs that were considered in the discovery set. Since the E2 individuals were not all genotyped on the same platform, we used imputed SNPs that passed the quality control filters when necessary; this resulted in 1,386 Neanderthal variants. For the non-Neanderthal GRM used in the additional two GRM replication analysis, we included all high quality non-Neanderthal variants with a MAF > 1% that were not within 100 kb of a Neanderthal variant (370,306 variants).

Phenotype Selection

We tested a manually curated set of PheWAS codes for ocular, brain, immune, lipid metabolism, digestive, or skin traits: myocardial infarction; depression; corns and callosities; mood disorders; overweight; seborrheic keratosis; coronary atherosclerosis; acute upper respiratory infections; obesity; anxiety disorder; hypercholesterolemia; actinic keratosis; other dermatoses; atopic or contact dermatitis; ischemic heart disease;

visual disturbances; diverticulosis and diverticulitis; diverticulosis; anxiety, phobic and dissociative disorders; hypermetropia; type 2 diabetes; peripheral arterial disease; hypovolemia; disorders of lipoid metabolism; hyperlipidemia; benign neoplasm of colon; astigmatism; disorders of fluid, electrolyte, and acid-base balance; symptoms affecting skin; glaucoma; Crohn's disease; irritable bowel syndrome; age-related macular degeneration; disorders of function of stomach; disorders of refraction and accommodation; retinal disorders; inflammation of the eye; other disorders of eye; macular degeneration; superficial cellulitis and abscess; electrolyte imbalance; cataract; inflammation of eyelids; hypertension; myopia; and disorders of vitreous body.

These categories were selected to represent traits that Neanderthal introgression has been hypothesized to influence in previous studies. Phenotypes tested in the discovery analyses either had a case prevalence > 20% or had an association with a nominally significant *P* value in a preliminary PheWAS analysis; 46 phenotypes met these criteria. The phenotypes tested in replication analyses had a *P* value < 0.1 in the discovery analyses (12 phenotypes).

Discovery and Replication Analyses

In the discovery analyses, we used GCTA to estimate the variance in risk explained by Neanderthal SNPs for 46 phenotypes using Neanderthal GRMs generated as described above. In the replication analyses, we tested the 12 phenotypes nominally significant in the discovery analysis using the E2 Neanderthal SNP GRM. We additionally tested these 12 replication phenotypes in a GCTA analysis with a Neanderthal and non-Neanderthal GRM fitted in the same model. We included age, sex, and eMERGE site as covariates in both replication and discovery analyses. In each analysis, we used disease prevalence estimates when available: Depression (15.0%) (*36*), Actinic keratosis (38.0%) (*37*), Obesity (30.2%) (*38*), Hypercholesterolemia (26.9%) (*39*), Anxiety disorder (18.0%) (*40*).

To ensure that the differences in percent risk estimated between E1 and E2 were not due to the variants that did not pass QC in E2, we also reran our discovery analyses without these variants. There was negligible difference between those results and our original results.

Best Linear Unbiased Predictions

Individual Neanderthal SNP effects were estimated using the best linear unbiased prediction (BLUP) approach in the GCTA package (*15*). We calculated BLUPs for the 12 significant or nominally significant phenotypes in both E1 and E2. We analyzed the genomic distribution of the 10% of SNPs with the highest and lowest BLUPs for actinic keratosis and depression using the Genomic Region Enrichment of Annotations Tool (GREAT) with the default basal plus extension settings (*41*).

Phenome-wide Association Scans

We performed both a meta-analysis of PheWASes on each eMERGE site's data individually and a joint PheWAS analysis over data pooled from all eMERGE sites. Both analyses were performed separately on the independent E1 (discovery) and E2 (replication) cohorts. We analyzed 1,495 common (MAF > 1%) Neanderthal SNPs genotyped by the eMERGE Network and required that phenotypes have at least 20 cases

in each site analyzed in the meta-analysis or overall for the pooled analysis. For the meta-analyses, a PheWAS was performed for each eMERGE site's data using the "phewas" function in the PheWAS package (*31*). A meta-analysis of the site-specific scans was performed with the "phewasMeta" function. We considered age, sex, and first three principal components as covariates. For the joint analyses, the "phewas" function in the PheWAS package was used to analyze data pooled across eMERGE sites. Covariates used were age, sex, eMERGE site, and the first three principal components. For imputed SNPs in the E2 analyses, we used dosages rather than the hard calls. We used gtool (v0.7.5) and qctool (v1.4) to select the appropriate SNPs from the IMPUTE2 files and convert to the input format for the PheWAS package. We report the *P* value and odds ratio from the fixed effect models unless otherwise stated.

Neanderthal Phenotype Association Enrichment Analyses

To explore whether Neanderthal SNPs were more likely to be associated with disease phenotypes than non-Neanderthal SNPs, we identified phenotype associations for 1,056 non-redundant ($r^2 < 0.5$) common (MAF > 1%) Neanderthal SNPs at a relaxed significance threshold of $P < 0.001$ in the discovery set that replicated in E2 ($P < 0.05$ and same direction of effect). This yielded 60 associations after accounting for hierarchically related phenotypes (Table S5). To generate a set of appropriate non-Neanderthal SNP-phenotype associations for comparison, we identified SNPs that were not within 100 kb of a Neanderthal SNP, and then we pruned these non-Neanderthal SNPs so that none had $r^2 > 0.5$. We then identified five independent matched control sets (for a total of 5,280 non-Neanderthal SNPs) that matched the allele frequency distribution (difference per matched SNP frequency < 0.005%) and genotyping status on the Human660W-Quadv1_A genotyping platform of the Neanderthal set. We performed a PheWAS meta-analysis of the non-Neanderthal SNPs following the same protocol as above using the hard calls for E2.

Supplementary Text

PheWAS Analysis

Pooled PheWAS Analysis

In addition to the PheWAS meta-analysis over eMERGE sites presented in the main text, we performed a pooled PheWAS analysis on combined E1 data from the different eMERGE sites and performed a replication PheWAS analysis on the pooled E2 data. The motivation for carrying out this additional pooled analysis was that it enabled us to test additional phenotypes that did not meet our inclusion criteria of at least 20 cases per site in the meta-analysis.

In general, we prefer the meta-analysis approach, because there are differences in the eMERGE individuals from different sites that could introduce biases when data are pooled that might not be fully accounted for by including site and principal components as covariates in the model. For example, clinicians within a site are more likely to use similar criteria for defining phenotypes and have similar billing practices than clinicians between sites. (Different hospitals may have different policies on the specific ICD-9 codes reported for a condition due in part to the local preferences of insurance companies.) Sites within eMERGE also ascertained their cohorts for different diseases, and these ascertainment differences could lead to biases in the phenotypes present in each set. In addition, there are geographic, environmental, and demographic (ancestry, age, gender) differences across sites that influence what diseases are likely to be present. Thus, we are most confident in an association when it is present across multiple eMERGE sites in our meta-analysis and it replicates in E1 and E2, but we report the pooled analysis here for completeness (Table S4).

The pooled PheWAS analysis found two locus-wise Bonferroni significant, replicated associations: rs12049593 with protein-calorie malnutrition and rs3917862 with hypercoagulable state. It also found five additional associations significant at $P < 10^{-4}$ in E1 that also replicated in E2 (Table S4). These included associations with gastroparesis, two sleep disorders, complications during pregnancy, and lung disease.

Comparison of Pooled and Meta Analysis

The two top Bonferroni significant associations in the meta-analysis were both found in the pooled analysis. The third significant replicating association in the meta-analysis, rs11030043 with symptoms of the urinary system, was significant in the E1 pooled analysis ($P = 1.3 \times 10^{-6}$), but the association was not significant in E2 ($p = 0.25$). It is possible that the diversity of the diseases that fall under this phenotype code lead to heterogeneity in its application between sites that may result in a less coherent case group when individuals are pooled across sites. The fourth significant replicating association in the meta-analysis, rs901033 with tobacco use disorder, is nominally significant in E1 ($P = 1.45 \times 10^{-4}$) and E2 ($P = 2.9 \times 10^{-3}$) in the pooled analysis.

Comparing the top nominally significant replicating associations between the meta (Table S3) and pooled analyses (Table S4), we found that associations with sleep related movement disorders and obstructive sleep apnea are present in both. Seven associations are unique to the meta-analysis and two are unique to the pooled analysis. As suggested above, there are many reasons why an association might have been detected in

one analysis and not the other, and lack of presence on both lists should not necessarily cast doubt on an association. For example, gastroparesis could not be tested in the E1 meta-analysis due to an insufficient number of cases in any individual site (54 cases overall and a maximum of 18 in a single site). Similarly, tests involving alveolar and parietoalveolar pneumopathy had reduced power in the meta-analysis compared to the pooled, because all 63 cases were considered in the pooled analysis, while only one site had a sufficient number of cases (21) to be considered in the meta-analysis. Conditions of mother complicating pregnancy did not reach the 10^{-4} nominal significance threshold in the meta-analysis (E1 $P = 2.3 \times 10^{-4}$); we suspect that this is likely due to the presence of fewer controls, as the number of cases was the same for both analyses.

Replication Values for PheWAS Meta-analysis

We also applied a new FDR-based method for evaluating whether focused follow-up studies replicate findings of a preliminary large-scale study (*42*) to the *P* values obtained from the E1 meta-analysis and the separate E2 meta-analysis. Using default parameters and thresholds from the main text, we found that the Neanderthal SNP associations described here (Table 2) have *r* values less than 0.05, except for the association with symptoms of the urinary system (Hypercoagulable state: 0.003; Protein-calorie malnutrition: 0.003; Tobacco use disorder: 0.017; Symptoms involving urinary system: 0.86). However, we note that the *r* value test correction does not fit our experimental design exactly, as we performed two complete independent meta-analyses on the discovery and replication cohorts. Ultimately, we believe that the biological relevance of the genomic regions containing the Neanderthal haplotypes to the associated phenotypes provides strong support for the associations.

BLUPs for Individual Neanderthal SNPs in GCTA Analyses

The GCTA analyses test the influence of a set of SNPs on the variance in a trait; some of the SNPs may increase risk while others may decrease risk. To explore if the Neanderthal SNPs were more or less likely to have a protective effect on the phenotypes found in the discovery GCTA analyses, we computed the best linear unbiased prediction (BLUP) for each Neanderthal SNP's effects. The proportion of risk SNPs for each trait in E1 and E2 are found in Table S1. The Neanderthal SNPs were not clearly shifted toward risk increasing or protective roles for the phenotypes examined. This is expected, since most variants will have little influence on most of the traits.

GREAT analysis of the genomic locations of Neanderthal SNPs with large magnitude BLUPs for actinic keratosis were enriched in regions annotated with many functions, including keratinocyte differentiation, several skin cancers, and many immune cell phenotypes (Table S2). The immune enrichment is striking given the role that modulation of the immune system plays in occurrence and treatment of actinic keratosis (*43*). The Neanderthal SNPs with large negative BLUPs for depression were enriched for basal ganglia disease, Parkinson's, and cranial nerve disease. The large positive BLUPs for depression were enriched for a diverse range of phenotypes including many cell migration and proliferation annotations, several neural phenotypes, and circadian clock genes (Table S2).

Neanderthal SNPs are associated with different classes of phenotypes than expected from non-Neanderthal SNP–phenotype associations.

We tested whether Neanderthal SNPs were more likely to be associated with disease phenotypes than non-Neanderthal SNPs. We compared the Neanderthal SNP PheWAS results to those obtained in a PheWAS of 5,280 SNPs with low LD ($r^2 < 0.5$) and an allele frequency distribution matched to the Neanderthal SNPs. These control SNPs correspond to five separate frequency-matched sets of SNPs. The Neanderthal SNPs were 1.22 times more likely to be associated with a phenotype than non-Neanderthal SNPs; however, due to the small number of associations these differences did not reach significance at $P < 0.05$.

To consider a larger number of associations, we analyzed all Neanderthal SNP–phenotype associations at a relaxed significance threshold of $P < 0.001$ in the discovery set that replicated ($P < 0.05$ and same direction of effect). This yielded 60 associations after accounting for hierarchically related phenotypes (Table S5). Of the 60 associations, 59 (98%) were risk increasing. We compared these results to the 260 associations obtained for the non-Neanderthal SNPs at the relaxed threshold. The Neanderthal SNPs were 1.12 times more likely to be associated with a phenotype than non-Neanderthal SNPs and were less likely to be protective (2% vs. 5%); however, these differences were not significant at $P < 0.05$ (binomial test, $P = 0.2$ and 0.13, respectively).

These results suggest that Neanderthal SNPs may be more likely to be associated with phenotypes compared to genotyped non-Neanderthal SNPs with the same allele frequency distribution; however, more data are needed to resolve this question. As more individuals are incorporated into EHR-linked genetic databases and additional whole-genome sequencing data become available for these individuals, it will be possible to more robustly test this hypothesis using our approach. Ultimately, the result of these analyses will provide insight into the genetic architectures of the traits influenced by admixture and the strength of purifying selection experienced by introgressed Neanderthal alleles.

Next, to test whether specific classes of phenotypes were more likely to be influenced by Neanderthal SNPs, we grouped PheWAS phenotypes into 14 distinct categories used in previous PheWAS studies and compared the distribution of associations for Neanderthal and non-Neanderthal SNPs. Overall, the Neanderthal SNPs influenced a significantly different distribution of phenotypes (chi squared test, $P = 0.017$; Figure 2). They were associated with more neurological (binomial test, $P = 0.018$) and psychiatric phenotypes ($P = 0.023$), and fewer digestive phenotypes ($P = 0.004$). These analyses suggest that Neanderthal alleles influence a different set of phenotypes than expected from non-Neanderthal alleles and may be more likely to contribute to disease.

To test if these enrichments and depletions were stable, we used the fact that the 5,280 control (non-Neanderthal) alleles consisted of five independent, non-overlapping sets matched to the Neanderthal alleles tested. We compared the Neanderthal phenotype association distribution to each of these five smaller matched sets in turn, and the phenotype categories at the extremes (psychiatric, neurological, and digestive) were all consistently enriched/depleted across the five comparisons. In particular, there was enrichment for psychiatric phenotype associations in the Neanderthal set across comparisons with all five sets (binomial test, $P < 0.05$). The enrichment for neurological

phenotypes was significant ($P < 0.05$) for three and trending ($P < 0.2$) for the remaining two. The depletion for digestive phenotypes was present in all five control set comparisons ($P < 0.05$). No other phenotypes were consistently enriched or depleted in more than two of the comparisons. Thus, our finding that Neanderthal alleles are associated with a significantly different set of traits than matched non-Neanderthal alleles is stable across different control sets, and the same phenotypes were consistently significantly enriched and depleted.

Neanderthal SNPs are enriched for brain eQTL

Given the observed enrichment for psychiatric and neurological phenotype associations among Neanderthal SNPs, we tested whether Neanderthal SNPs were more likely to be expression quantitative trait loci (eQTL) in brain tissues than non-Neanderthal SNPs. We analyzed previously computed brain eQTL datasets from cerebellum and temporal cortex from Zou *et al.* (2012) (*44*) and cerebellum and parietal cortex from ScanDB (*45*).

Zou et al. quantified expression levels of 24,526 transcripts in the cerebellum and temporal cortex of autopsied patients with Alzheimer's disease (AD; 197 cerebellum, 202 temporal cortex) and patients with other brain pathologies (non-AD; 177 cerebellum, 197 temporal cortex) using Illumina's Whole Genome DASL assay. The patients were genotyped on the Illumina HumanHap300-Duo Genotyping BeadChip. They then tested SNPs within 100 kb of the quantified transcripts for association with expression level. These analyses were performed for the AD, non-AD, and combined cohorts for each tissue. To maximize power, we analyzed the association *P* values from the combined set.

We identified all Neanderthal and control non-Neanderthal SNPs directly genotyped and tested in this study that were not overlapping a probe on the array. This yielded 663 Neanderthal SNP–gene pairs with association *P* values and 3,295 non-Neanderthal SNP–gene pairs with *P* values. To correct for multiple testing, we calculated q-values (*46*) from the raw Neanderthal and non-Neanderthal SNP-gene pair *P* values. At a q-value threshold of 0.05, 22 of the 663 Neanderthal SNP–gene pairs (3.3%) and 45 of 3,295 non–Neanderthal SNP-gene pairs (1.4%) were significant eQTL in the cerebellum. This enrichment of eQTL among the Neanderthal SNPs is significant ($P = 1.68E-04$, one-tailed binomial test). We also found significant enrichment for temporal cortex eQTL among the Neanderthal SNPs: 23 of 683 Neanderthal SNP-gene pairs (3.4%) and 42 of 3,298 non-Neanderthal SNP-gene pairs (1.3%) were eQTL ($P = 3.49E-05$, one-tailed binomial test). These results were robust to q-value thresholds of 0.01 and 0.1.

We repeated the enrichment analysis using only unique SNPs from the significant SNP–gene pairs above. These comparisons also revealed significant enrichment for brain eQTL among Neanderthal SNPs: 21 of 297 Neanderthal variants (7.1%) and 44 of 1,462 non-Neanderthal variants (3.0%) were found in the cerebellum ($P = 3.2E-04$, one-tailed binomial test). In the temporal cortex, 19 of 307 Neanderthal variants (6.2%) and 42 of 1,482 non-Neanderthal variants (2.8%) were eQTL for at least one gene ($P = 1.4E-03$). In all, 29 unique Neanderthal SNPs were brain eQTL for at least one transcript in the cerebellum or temporal cortex.

We also analyzed brain eQTL in the cerebellum and parietal cortex from the ScanDB database computed from expression and genotyping data originally collected by

the Bipolar Disorder Genome Study (BiGS) Consortium (*47*). ScanDB provides only the subset of the SNP–gene expression association *P* values for tests with an uncorrected $P < 0.01$. We analyzed all pairs they defined as significant by this threshold and identified whether each variant acted as an eQTL for any tested gene. In the cerebellum, 168 of 1,056 Neanderthal variants (15.9%) and 734 of 5,280 non-Neanderthal variants (13.9%) were nominal eQTL; this represents significant enrichment ($P = 0.035$, one-tailed binomial test). However, in the parietal cortex, 158 Neanderthal variants (15.0%) and 742 non-Neanderthal variants (14.1%) acted as eQTL for at least one gene. This difference was not significant at the 0.05 level ($P = 0.209$, one-tailed binomial test).

In summary, we find that Neanderthal SNPs are significantly enriched for eQTL activity in the cerebellum and temporal cortex. The enrichment for cerebellar eQTL activity replicated in an independent cohort. Our results suggest that further dissection of the contribution of Neanderthal alleles to gene expression patterns in the human brain holds promise.

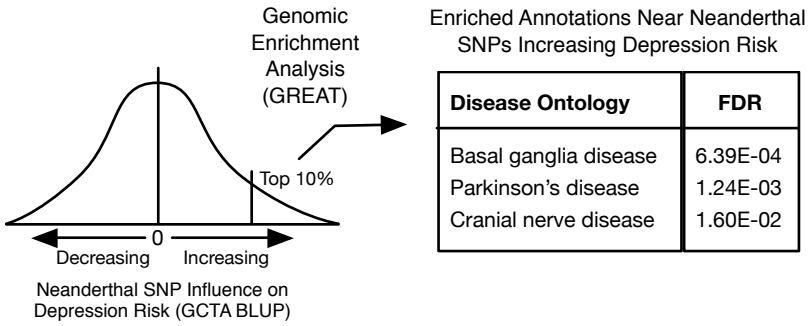
The relationship between rs3917862 and the Factor V Leiden Mutation

In our PheWAS, the non-coding Neanderthal SNP rs3917862 was significantly associated with hypercoagulable state in E1, and this association replicated in E2 (Table 2; Figure 1D). This SNP had functional genomics marks suggestive of gene regulatory activity (Figure S2), and we found that it associated with significantly increased expression of *SELP* and *F5* in arteries (Figure 1E; Figure S3). These findings support its association with hypercoagulability, because increased levels of *SELP* and *F5* both increase risk for diseases linked to hypercoagulability, such as deep vein thrombosis, embolism, and (indirectly) miscarriage (*48, 49*). Indeed, rs3917862 has independently been associated with venous thromboembolism (VTE) (*50*).

However, due to the large odds ratio (~3) for the association and the proximity (~74 kb downstream) of rs3917862 to the *F5* Leiden thrombophilia mutation (F5L, rs6025), which increases risk for several conditions linked to hypercoagulability in individuals of European ancestry, we investigated whether this Neanderthal SNP could tag associations due to F5L. F5L is overlapped by a Neanderthal haplotype, but appears to postdate introgression. It was not genotyped on the arrays used by eMERGE, but we found modest linkage disequilibrium (r^2=0.07, D'=0.42) with the imputed F5L and rs3917862. This is in agreement with previous studies ($r^2 = 0.12$, D' = 0.37) (*21*) and our analysis of sequencing data from 1000 Genomes Phase 3 EUR individuals ($r^2 = 0.06$, D' = 0.56). Furthermore, manual review of the EHRs for hypercoagulable state cases revealed that only four had a positive F5L genetic test out of 11 directly tested. Using the imputed F5L data, we tested whether we had power to detect an association with hypercoagulable state caused by F5L via rs3917862. We took the imputed frequency of the F5L mutation (2.9%). We used estimates for the genotype relative risk (Aa = 10, AA = 20) from odds ratio estimates of the association with imputed F5L with hypercoagulable state. We took the frequency of rs3917862 (6.2%), hypercoagulable state prevalence (1.6%), case numbers (92), and control numbers (9,540) from the E1 data to compute the power of rs3917862 to tag the F5L association. At our Bonferroni-corrected alpha threshold, we were significantly underpowered to detect an association driven by F5L via rs3917862 (dominant model: 36%; allelic model: 39%) (*51*). We also tested a range of values that

reflected the extremes of the estimates of these values from the literature. Nearly all remained significantly underpowered; however in a few situations, increasing the F5L mutation frequency to ~5% yielded power above 80%.

It is possible that the F5L mutation contributes to the significant observed association between hypercoagulable state and rs3917862. However, the modest LD between these SNPs, the lack of positive F5L tests in the reviewed cases, and our evidence that rs3917862 influences the expression of *F5* and *SELP* in a manner consistent with increased risk suggests an additional role for the Neanderthal allele in hypercoagulability. Furthermore, a recent well-powered study of VTE demonstrated that rs3917862 increases the risk of VTE beyond the risk associated with F5L (*21*). Thus, we conclude that this Neanderthal allele influences hypercoagulable state.



Enriched Annotations Near Neanderthal SNPs Increasing Depression Risk

Disease Ontology	FDR
Basal ganglia disease	6.39E-04
Parkinson's disease	1.24E-03
Cranial nerve disease	1.60E-02

Fig. S1. Schematic example of functional enrichment analysis on genes nearby Neanderthal SNPs with large BLUPs in the GCTA analyses.

We estimated the effects of individual Neanderthal SNPs (BLUPs) and performed genomic enrichment analysis using GREAT (*41*) on the top 10% most protective and risk increasing SNPs for actinic keratosis and depression. We found enrichment (FDR < 0.05; hypergeometric test) for many functional annotations: most notably, keratinocyte differentiation and several immune functions for actinic keratosis and regions involved in neurological diseases, cell migration, and circadian clock genes for depression (Table S2)

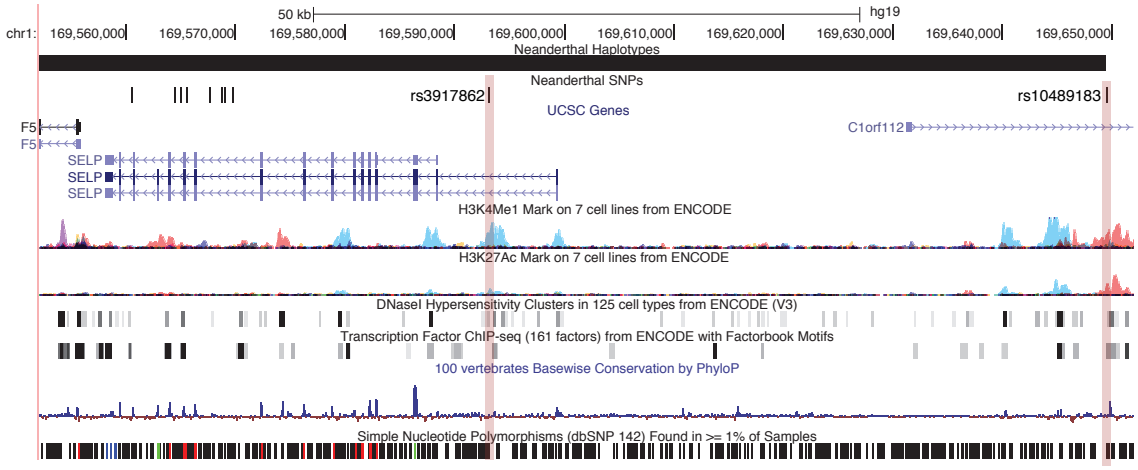


Fig. S2. A Neanderthal allele significantly associated with hypercoagulable state and located in *SELP* and has evidence of gene regulatory function.

The Neanderthal SNP, rs3917862, is significantly associated with hypercoagulable state (Figure 1D; Table 2). Rs3917862 is located in an intron of P-selectin (*SELP*), a gene that mediates leukocyte action at injuries in the early stages of inflammation. This SNP is in LD with rs10489183; these SNPs have functional genomic signatures indicative of gene regulatory activity in blood cells and vein epithelial cells. The Neanderthal allele at rs3917862 is significantly associated ($P = 0.005$) with increased expression of *SELP* in tibial artery data from GTEx (Figure 1E). It also significantly associated with increased *F5* expression ($P = 0.05$; Figure S3).

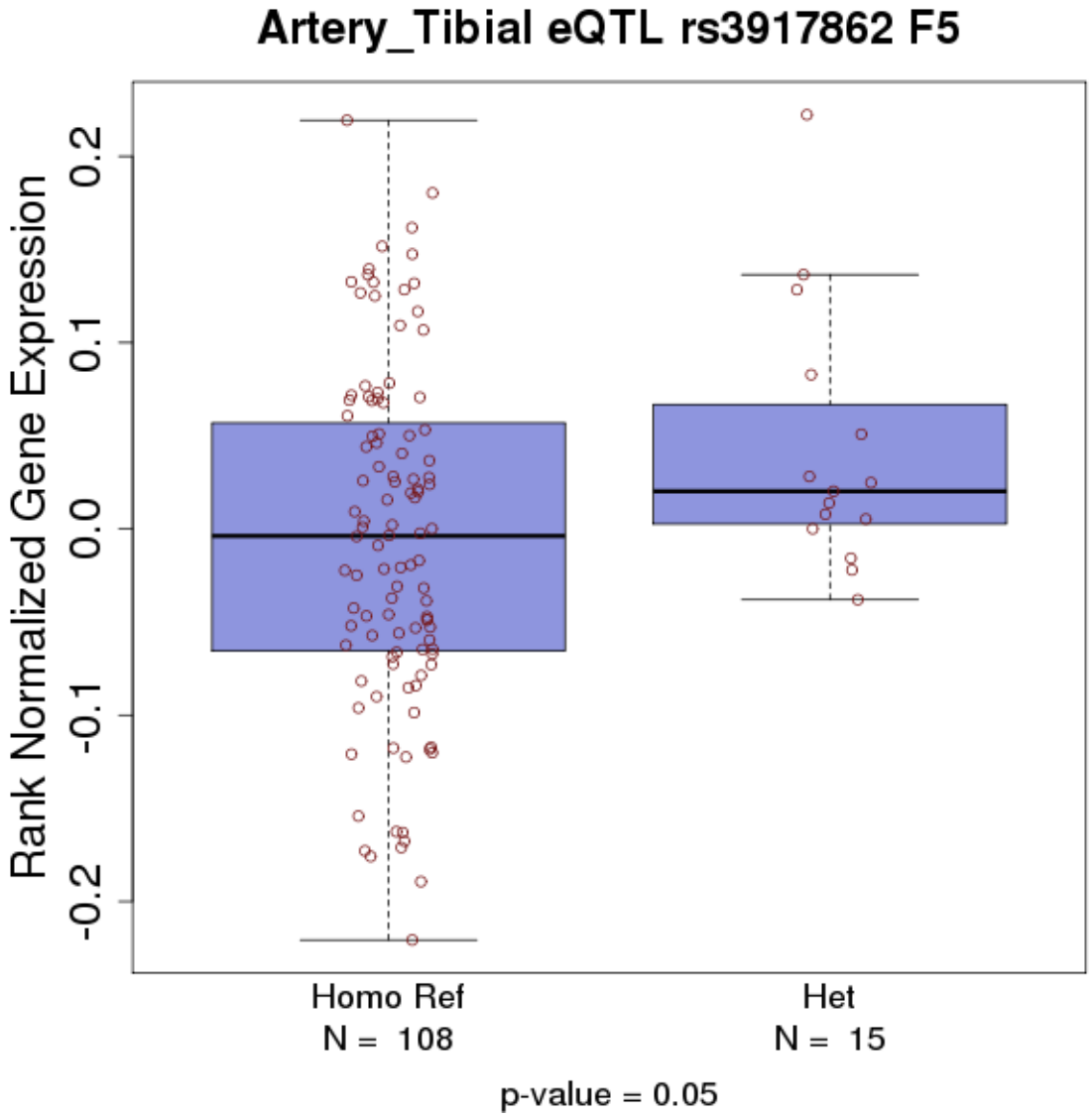


Fig. S3. rs3917862 is significantly associated with increased expression of F5 in tibial artery (GTEx).

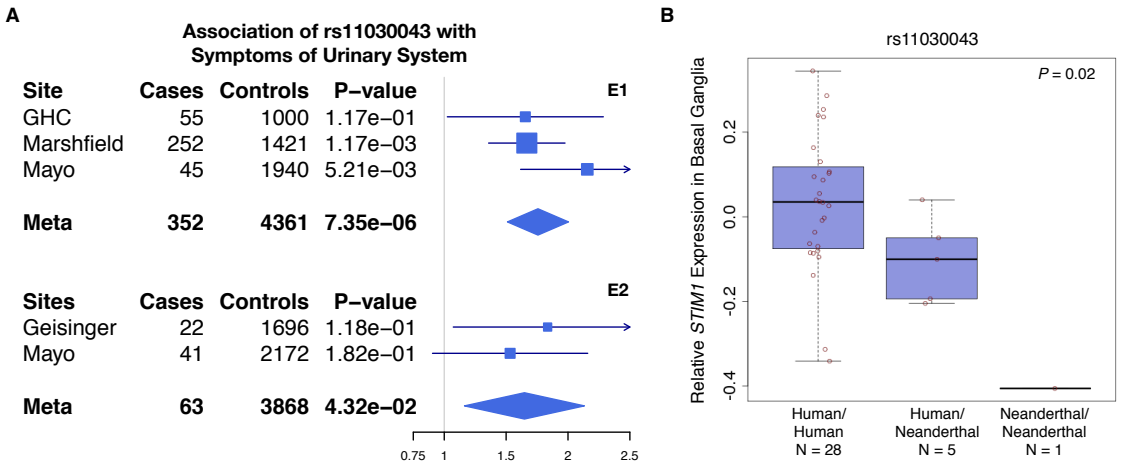


Fig. S4. rs11030043 is significantly associated with symptoms of the urinary system, and is associated with expression levels of *STIM1*.

(A) The forest plot shows odds ratios and *P* values for the association of Neanderthal SNP rs11030043 with symptoms of the urinary system in each site with >= 20 cases. This association was significant in E1 and replicated in E2. **(B)** rs11030043 is located ~10 kb upstream of stromal interaction molecule 1 (*STIM1*), a transmembrane protein that regulates calcium ion flux. In GTEx samples, the Neanderthal allele is significantly associated ($P = 0.02$) with increased expression of *STIM1* in caudate basal ganglia, a region of the brain connected to bladder dysfunction, particularly in those with neurological conditions such as Parkinson's.

Table S1. Neanderthal SNP BLUP results from GCTA.

Significantly replicating results are in bold. PRE = Percent Risk Explained.

Phenotype	E1 PRE	E1 % Risk SNPs	E2 PRE	E2 % Risk SNPs
Hypercholesterolemia	0.74%	49.4%	0.20%	50.1%
Overweight	0.60%	46.6%	0.23%	48.6%
Obesity	0.59%	48.1%	0.39%	48.0%
Mood disorders	**1.11%**	**48.2%**	**0.68%**	**52.0%**
Depression	**2.03%**	**48.1%**	**1.06%**	**52.4%**
Anxiety disorder	1.70%	50.7%	0.00%	49.2%
Myocardial Infarction	1.39%	50.6%	0.13%	50.1%
Coronary atherosclerosis	0.68%	51.0%	0.34%	53.4%
Acute upper respiratory infections	0.70%	47.5%	0.34%	49.8%
Corns and callosities	1.26%	48.6%	0.21%	50.5%
Actinic keratosis	**0.64%**	**49.7%**	**2.49%**	**54.1%**
Seborrheic keratosis	0.77%	50.0%	0.41%	52.2%

Table S2. GREAT analysis of Neanderthal SNPs with highest/lowest BLUPs from actinic keratosis and depression GCTA tests.

Excel spreadsheet available online.

Table S3. Nominally significant (discovery $P < 10^{-4}$) replicating results from meta-analyses.

Significant results (from Table 2) are in bold. Results shaded in light grey were not found in joint analysis (discovery $P < 0.001$, replication $P < 0.05$, consistent OR).

Phenotype	SNP	Flanking Gene(s)	Discovery		Replication	
			Odds Ratio	*P*	Odds Ratio	*P*
Hypercoagulable state	**rs3917862**	**SELP**	**3.32**	**9.9E-07**	**3.00**	**5.0E-10**
Protein-calorie malnutrition	**rs12049593**	**SLC35F3**	**1.77**	**2.0E-06**	**1.63**	**5.5E-05**
Symptoms involving urinary system	**rs11030043**	**RHOG, STIM1**	**1.76**	**7.4E-06**	**1.65**	**4.3E-02**
Tobacco use disorder	**rs901033**	**SLC6A11**	**2.19**	**1.7E-05**	**1.75**	**7.9E-04**
Hemangioma and lymphangioma, any site	rs17114127	PPAP2B	2.61	4.5E-05	3.41	4.3E-04
Functional disorders of bladder	rs17115796	DAB1	2.40	5.4E-05	1.85	2.2E-02
Stress incontinence, female	rs17766531	PRDM15	1.53	5.4E-05	1.34	3.5E-02
Microscopic hematuria	rs35609966	CCR6, GPR31	2.51	6.5E-05	2.81	3.6E-02
Obstructive sleep apnea	rs7133666	PIK3C2G	1.36	7.2E-05	1.22	8.7E-03
Malignant neoplasm of brain and nervous system	rs3783796	PRKCH	5.63	8.2E-05	4.04	9.6E-03
Stiffness of joint	rs11817964	ZNF365	1.94	8.7E-05	1.62	5.0E-02
Psoriasis vulgaris	rs12190231	EEF1E1, SLC35B3	1.48	9.1E-05	1.31	1.4E-02
Sleep related movement disorders	rs3771635	PKP4	0.70	9.8E-05	0.75	3.9E-03

Table S4. Nominally significant (discovery $P < 10^{-4}$) replicating results from joint analyses.

Results reaching locus-wise Bonferroni corrected threshold are in bold. Results shaded in light grey were not found in the meta-analysis (discovery $P < 10^{-4}$, replication $P < 0.05$, consistent OR).

Phenotype	SNP	Flanking Gene(s)	Discovery		Replication	
			Odds Ratio	*P*	Odds Ratio	*P*
Protein-calorie malnutrition	**rs12049593**	**SLC35F3**	**1.72**	**3.14E-06**	**1.52**	**3.56E-04**
Hypercoagulable state	**rs3917862**	**SELP**	**2.66**	**1.46E-05**	**2.67**	**3.66E-10**
Gastroparesis	rs4963700	SOX5	2.63	4.01E-05	1.48	2.25E-02
Sleep related movement disorders	rs3771635	PKP4	0.69	5.77E-05	0.73	1.28E-03
Obstructive sleep apnea	rs7133666	PIK3C2G	1.36	7.67E-05	1.20	1.68E-02
Other conditions of the mother complicating pregnancy	rs16868879	NCALD	5.39	7.81E-05	2.11	1.44E-03
Other alveolar and parietoalveolar pneumonopathy	rs10456309	KIAA0319	3.19	7.98E-05	1.81	3.57E-02

Table S5. Neanderthal SNP-phenotype associations used in the comparison with non-Neanderthal SNP-phenotype associations.

Redundant SNP-phenotype associations due to one SNP associating with multiple phenotypes in the same phenotype hierarchy are highlighted in gray.

Phenotype	SNP	Category	Discovery		Replication	
			Odds Ratio	*P*	Odds Ratio	*P*
Chronic airway obstruction	rs2300659	Pulmonary	1.24	9.92E-04	1.18	2.09E-02
Chronic pain syndrome	rs2298146	Neurologic	4.02	2.80E-04	2.74	1.88E-02
Hemangioma and lymphangioma, any site	rs17114127	Neoplastic	2.61	4.47E-05	3.41	4.28E-04
Functional disorders of bladder	rs17115796	Genitourinary	2.39	5.38E-05	1.85	2.20E-02
Calculus of ureter	rs12563768	Genitourinary	2.15	8.62E-04	2.08	6.59E-03
Coagulation defects	rs3917862	Hematopoietic	1.27	8.34E-04	1.24	1.05E-03
Hypercoagulable state	rs3917862	Hematopoietic	3.32	9.87E-07	3.00	5.00E-10
Skin neoplasm of uncertain behavior	rs16848353	Neoplastic	1.50	4.73E-04	1.31	8.52E-03
Other disorders of soft tissues	rs17675600	Musculoskeletal	2.99	3.38E-04	1.71	2.13E-02
Protein-calorie malnutrition	rs12049593	Endocrine & Metabolic	1.77	1.98E-06	1.63	5.46E-05
Hepatic cancer	rs17018123	Neoplastic	4.32	5.31E-04	2.73	3.68E-03
Sleep related movement disorders	rs3771635	Neurologic	0.70	9.78E-05	0.75	3.92E-03
Personality disorders	rs2288187	Psychiatric	2.21	7.94E-04	2.62	7.44E-03
Radiotherapy	rs901033	Neoplastic	2.55	5.67E-04	3.53	9.60E-03
Tobacco use disorder	rs901033	Psychiatric	2.19	1.73E-05	1.75	7.93E-04
First degree AV block	rs901033	Cardiovascular	3.21	2.58E-04	1.92	4.74E-02

Functional disorders of bladder	rs13087234	Genitourinary	1.61	7.49E-04	1.34	3.88E-02
Infections involving bone	rs17029555	Musculoskeletal	1.62	8.49E-04	1.49	4.24E-03
Rheumatoid arthritis & related inflammatory polyarthropathies	rs12639456	Musculoskeletal	1.68	5.00E-04	1.54	1.07E-02
Rheumatoid arthritis	rs12639456	Musculoskeletal	1.76	6.58E-04	1.65	5.83E-03
Acquired foot deformities	rs1242069	Musculoskeletal	1.42	5.05E-04	1.39	1.48E-02
Gram negative septicemia	rs2050807	Infectious	4.28	2.45E-04	2.48	3.82E-02
Atherosclerosis of aorta	rs13151936	Cardiovascular	1.45	9.24E-04	1.42	4.00E-02
Leukemia	rs17527711	Neoplastic	1.78	6.90E-04	1.37	3.45E-02
Age-related osteoporosis	rs10516526	Musculoskeletal	3.56	1.08E-04	1.63	9.44E-03
Acquired toe deformities	rs10517630	Musculoskeletal	1.52	9.97E-04	1.56	1.33E-02
Psoriasis & related disorders	rs12190231	Dermatologic	1.35	9.58E-04	1.29	1.18E-02
Psoriasis	rs12190231	Dermatologic	1.40	4.56E-04	1.32	8.39E-03
Psoriasis vulgaris	rs12190231	Dermatologic	1.48	9.14E-05	1.31	1.40E-02
Dry eyes	rs12189640	Neurologic	1.37	5.62E-04	1.31	3.16E-02
Disorders of other cranial nerves	rs12662332	Neurologic	1.71	6.59E-04	2.02	1.50E-04
Subjective visual disturbances	rs1513498	Neurologic	1.53	1.37E-04	1.67	1.40E-02
Other cerebral degenerations	rs3822947	Neurologic	2.00	3.24E-04	1.47	4.87E-02
Microscopic hematuria	rs35609966	Genitourinary	2.51	6.49E-05	2.81	3.62E-02
Cancer, suspected or other	rs9366117	Neoplastic	2.38	5.15E-04	1.88	2.94E-02
Proteinuria	rs17722435	Endocrine & Metabolic	2.83	2.02E-04	1.67	4.10E-02
Other conditions of the mother complicating	rs16868879	Genitourinary	5.89	2.34E-04	2.25	9.65E-04

pregnancy						
Memory loss	rs16914252	Psychiatric	1.77	6.23E-04	1.70	1.52E-02
Antisocial/borderline personality disorder	rs11139709	Psychiatric	4.92	5.28E-04	4.65	2.44E-03
Chronic kidney disease, Stage IV (severe)	rs1542479	Genitourinary	1.98	7.19E-04	1.63	2.04E-02
Polyp of female genital organs	rs17742994	Genitourinary	1.63	5.08E-04	1.46	2.51E-02
Stiffness of joint	rs11817964	Musculoskeletal	1.94	8.71E-05	1.62	5.00E-02
Allergies, other	rs2394616	Injuries	2.13	3.11E-04	1.78	2.51E-02
Inflammatory diseases of female pelvic organs	rs1931553	Genitourinary	1.57	8.70E-04	1.36	3.67E-02
Conduct disorders	rs12768228	Psychiatric	2.53	8.96E-04	2.31	4.26E-02
Symptoms involving urinary system	rs11030043	Genitourinary	1.76	7.35E-06	1.65	4.32E-02
Disorders of cornea	rs16905974	Neurologic	2.69	6.12E-04	2.69	1.07E-02
Obstructive sleep apnea	rs7133666	Neurologic	1.36	7.18E-05	1.22	8.66E-03
Erythematous conditions	rs17191680	Dermatologic	1.31	7.60E-04	1.25	8.98E-03
Emphysema	rs12579609	Pulmonary	1.88	9.30E-04	1.45	4.77E-02
Other conditions of brain	rs11060784	Neurologic	2.21	9.59E-04	1.55	3.77E-02
Neoplasm of unspecified nature of digestive system	rs9316483	Neoplastic	1.84	2.58E-04	1.97	2.27E-02
Malunion fracture	rs17080490	Musculoskeletal	3.91	4.52E-04	2.22	2.18E-02
Disorders of other cranial nerves	rs12896790	Neurologic	1.72	1.55E-04	1.58	1.05E-02
Malignant neoplasm of brain and nervous system	rs3783796	Neoplastic	5.63	8.19E-05	4.04	9.57E-03
Bipolar	rs11159544	Psychiatric	1.45	5.14E-04	1.31	2.79E-02
Fracture of humerus	rs4617810	Injuries	2.77	8.77E-04	1.86	1.27E-02

Chronic obstructive asthma	rs2240903	Pulmonary	2.75	3.35E-04	2.89	3.77E-03
Alopecia	rs17765170	Dermatologic	2.21	9.55E-04	2.86	1.43E-03
Generalized anxiety disorder	rs6122806	Psychiatric	3.23	9.34E-04	2.73	1.39E-02
stress incontinence, female	rs17766531	Genitourinary	2.95	6.94E-04	2.10	1.02E-02
Atherosclerosis of renal artery	rs5756326	Cardiovascular	1.53	5.38E-05	1.34	3.50E-02
Nontoxic multinodular goiter	rs2886122	Endocrine & Metabolic	1.45	5.14E-04	1.31	2.79E-02
Disorders of the autonomic nervous system	rs2281117	Neurologic	2.13	7.69E-04	2.05	4.28E-03